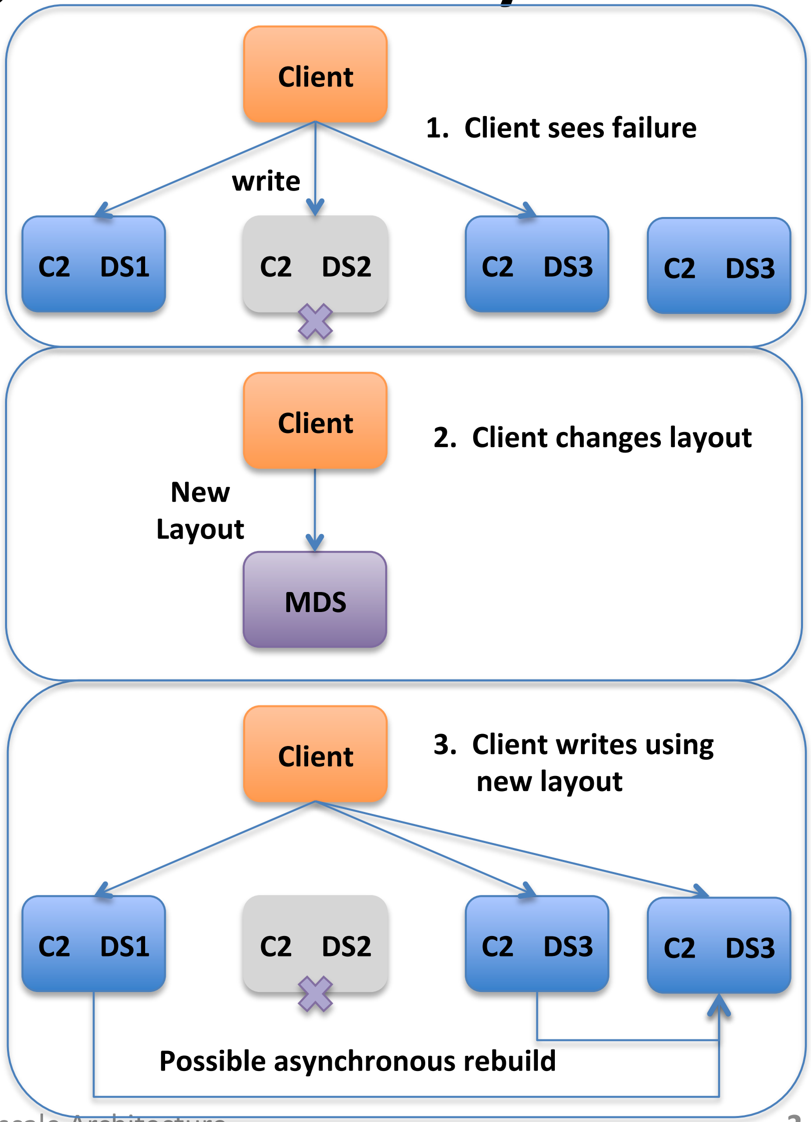
A few thoughts about future file systems

Peter Braam

HECFS 2010

Non blocking availability

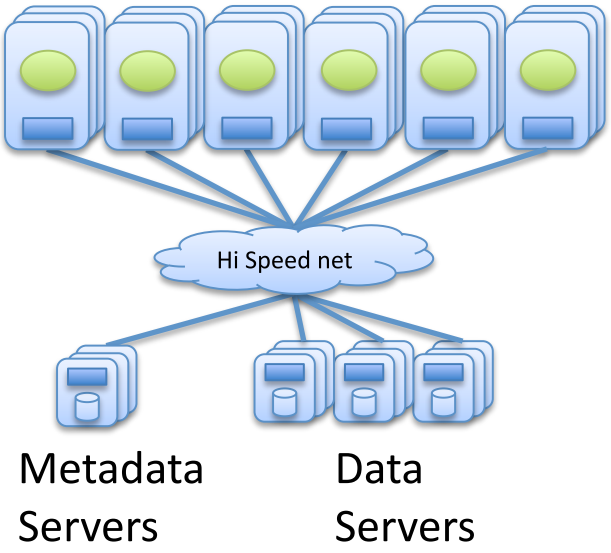
- Failures & overload: common
- Failover
  - Wait for resource
  - Doesn't work well
- Focus on availability
  - No reply (failure, load)
  - Adapt layout
  - Asynchronous cleanup
- Client determines timeout



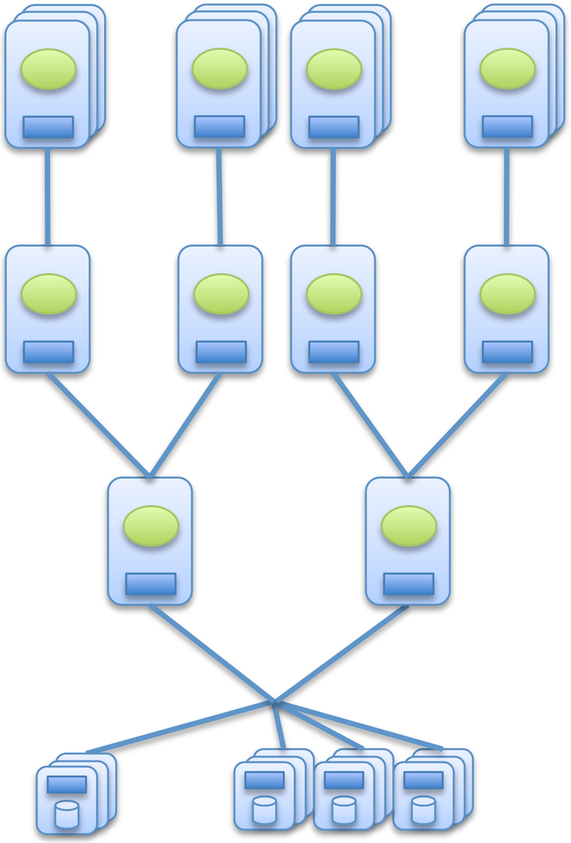
Data Layout ("striping") & Metadata

- Almost all layouts ("striping") will be regular
  - Use formulas, do not enumerate objects & devices
  - Use references between metadata tables
    - Avoid multiple copies
    - Results in very small MD probably < 150B/inode
- After failures, data layout becomes complex
  - Failures can move millions of different extents in a file
  - Maybe clean this up asynchronously
  - Are there good formulas for this?

Scalable communications



Physical Organization of Cluster



Logical Organization for Resource Management

Oh, you need to read?

- Flash cache doesn't help much for reading
- Physics – disks are slow
- Two common cases:
  1. Everyone reads the same – bit-torrent ideas
  2. Everyone reads something different – pre-stage

- Case 1 is addressed with scalable comms
- Case 2 can leverage log to predict pre-stage